CONSTELLATION

Use this tool to explore the relationships between genes in a gene list generated by differential expression, genome-wide knockout, or biomarker analyses

dependency map Consortium

WHAT

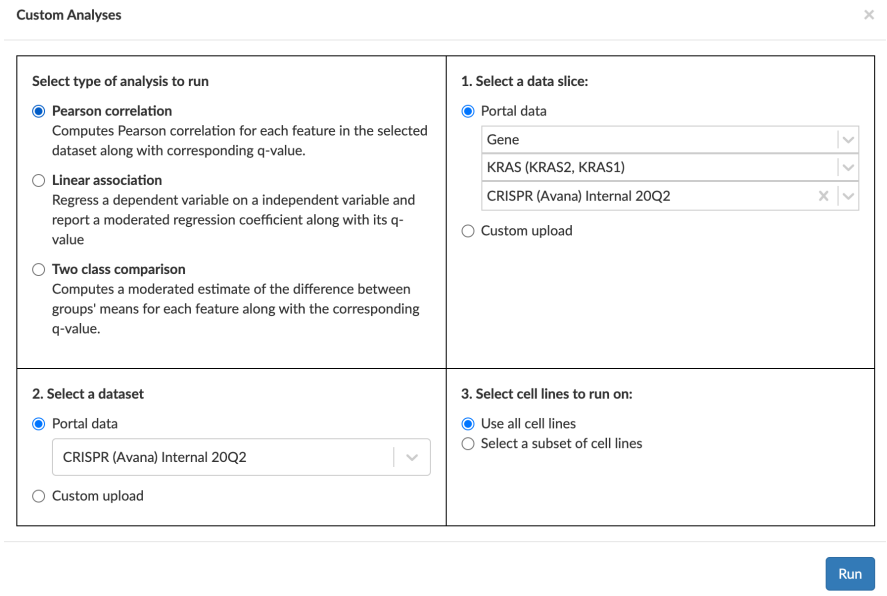
To explore KRAS co-dependencies

WHY

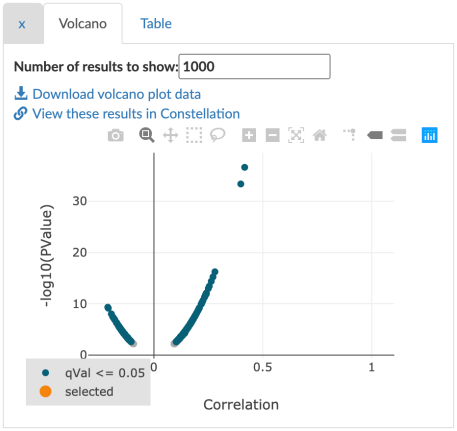
To learn about the function of KRAS

HOW

1. Use "Custom Analyses" in Data Explorer to correlate the KRAS Avana profile with all other Avana profiles



2. Click "View these results in Constellation"Or download volcano plot data from Custom Analysis and open the CSV in Constellation



BROAD INSTITUTE

dependency map Consortium

THINGS TO NOTE

- The KRAS codependencies form distinct clusters related to their function. For example, all of the MAPK genes cluster together.
- ERBB1 downstream pathway is the top gene set. ERBB1 is upstream of KRAS. If you select the ERBB1 downstream pathway a number of the top codependency are highlighted.
- Negative and positive codependencies can cluster together. For example, NRAS and RAF are connected.
- NRAS is a negative codependency of KRAS because activating mutations in NRAS and KRAS are generally mutually exclusive because both are upstream of RAF.

dependency map Consortium

WHAT

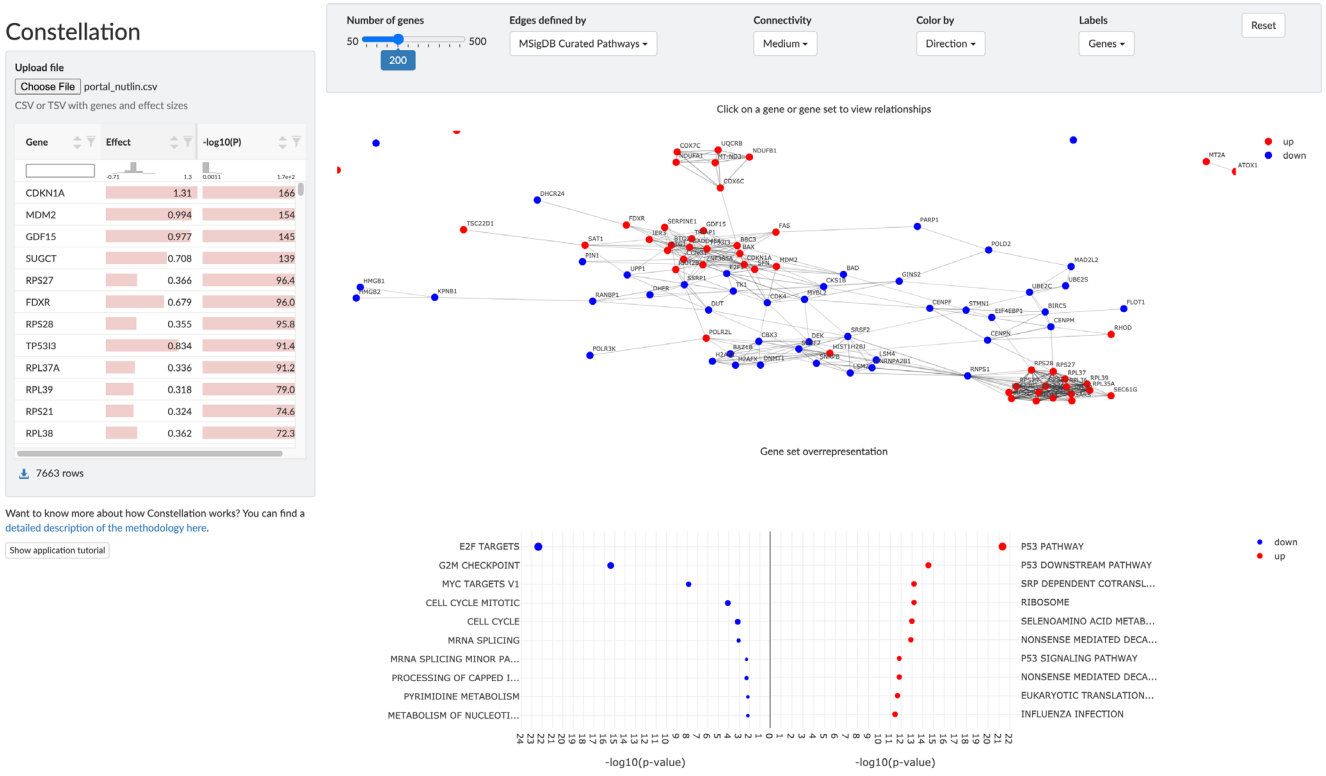
To understand your differential expression results

WHY

To learn about the function of the top differentially express genes

HOW

Load the results of your differential expression analysis into Constellation then select edges defined by MSigDB Curated Pathways

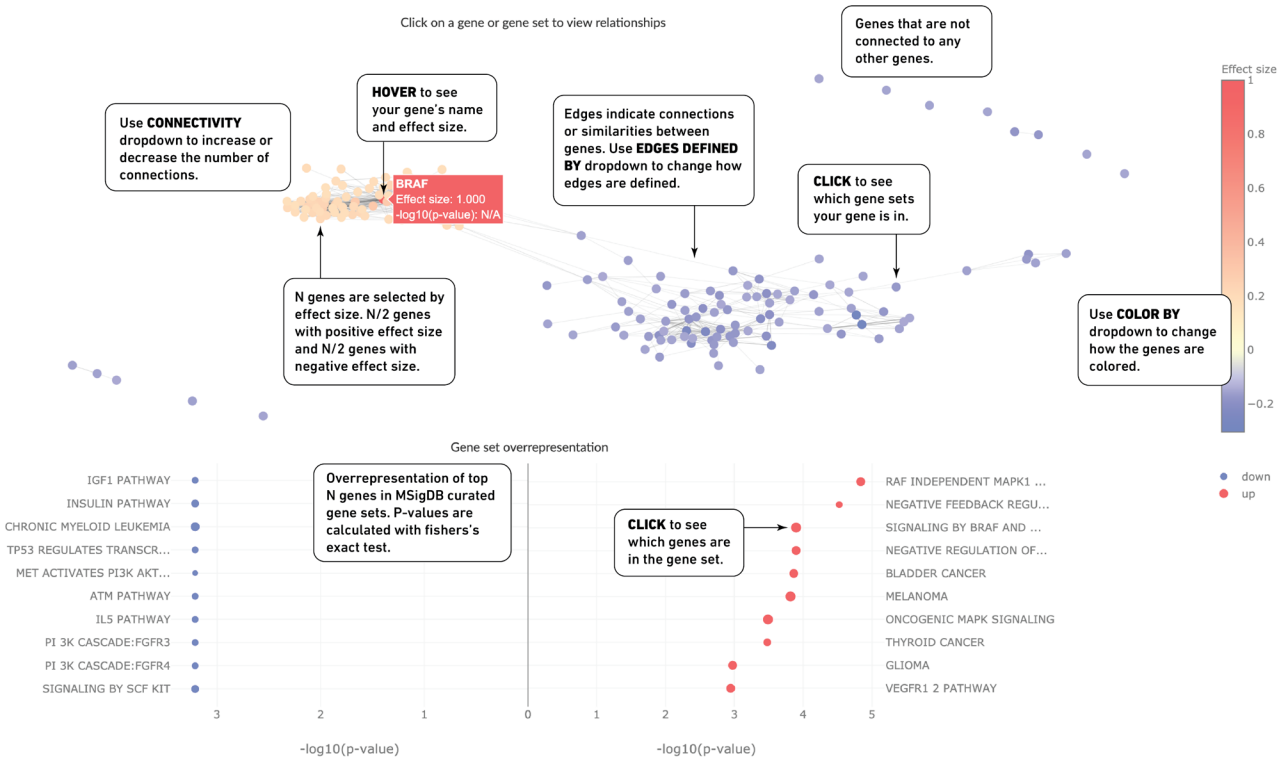


THINGS TO NOTE

- This example uses differential expression results from cancer cell lines treated with the MDM2 inhibitor Nutlin.
- When edges are defined by MSigDB Curated Pathways the distance between two genes is based on the similarity of their membership in MSigDB Curated Pathway gene sets.
- Genes in the same gene set tend to cluster together. For example, there is a large cluster of genes in the P53 Pathway gene set.

dependency map Consortium

OVERVIEW OF TOOL



METHODOLOGY

SELECTING GENES

When a CSV or TSV file is input the top N (set using the slider) genes are selected based on effect size. If the file contains both positive and negative effect sizes, N/2 genes with the greatest positive effect size and N/2 genes with the greatest negative effect size are selected. If the file only contains only positive or only negative effect sizes, N genes with the greatest effect size are selected.

EDGE TYPES

A number of different datasets can be used to define edges.

MSigDB curated pathways

An edge connecting two genes has a higher weight the more similar the two genes are to each other. Intuitively, two genes are similar if they are each members of the same, or similar, curated gene sets. Specifically, each gene is represented by a 64-dimensional vector learned from the gene set membership data of MSigDB's C2 Canonical Pathways (v7.0) and Hallmark (v7.1) gene set collections. StarSpace was used to learn the representations. Cosine distance was used to define the similarity between two gene representations.

STRING DB

Three separate STRING DB edge types are used: protein-protein interaction (PPi), literature, and combined. Edge weights come directly from the String-DB Homo sapiens v11 protein.links. detailed file.

Achilles codependency

Achilles codependency is calculated based on the DepMap internal 20Q2 achilles_gene_effect file. Edge weights are the Pearson correlation coefficient between gene pairs.

CCLE coexpression

CCLE coexpression is calculated based on the DepMap internal 20Q2 ccle_expression file. Edge weights are the Pearson correlation coefficient between gene pairs.

CONNECTIVITY

The connectivity setting specifies the size of the edge database used to connect the top N genes. High, medium, and low connectivity mean the same thing for all of the edge types:

- High: top 1% by weight of possible gene-gene edges are used.
- Medium: top 0.2% by weight of possible gene-gene edges are used.
- Low: top 0.004% by weight of possible gene-gene edges are used.

The number of possible gene-gene edges is set at 19321^2 since there are 19321 named protein-coding genes in the HGNC database. Since connectivity is set globally, the number of edges will vary between gene lists, and it is possible that some gene lists will have no edges.

NETWORK LAYOUT

The x and y position of each gene in the network is calculated using the Fruchterman-Reingold force-directed layout algorithm. Specifically, an undirected graph with weighted edges specified by the selected edge type and connectivity is input into the layout_with_fr function from the igraph R package. Genes that are not connected to any other genes are not laid out, instead, they are positioned at the bottom in a line ordered by effect size.

GENE SET OVERREPRESENTATION

P-values for the overrepresentation of the top N genes in the MSigDB v7.1 hallmark (H) and curated pathway (C2) gene sets are calculated with Fisher's exact test. The gene sets labeled positive are overrepresented in the N/2 genes with positive effect size and gene sets labeled negative are overrepresented in the N/2 genes with negative effect size.